랩 위드의 AI 기반 솔루션 >

"성능 1위"가 "도입 가능 1위"는 아니었습니다 — 국산 LLM 8종을 직접 실사해본 이야기

기술사업화 랩 위드
2026.07.20. 18:13 조회 10

기술사업화 랩 위드㈜는 AI 시스템을 납품하는 회사가 아닙니다.

저희 본업은 기업과 기관의 기술을 냉정하게 뜯어보고, "이 기술이 실제로 돈이 되는가"를 가려내는 일입니다.

기술이전, R&D 사전기획, 사전 경제성 평가 — 결국 모두 같은 질문으로 수렴합니다.

그런데 최근 저희가 사내 R&D 문서 작성에 쓸 sLLM(소형언어모델)을 고르면서, 평소 고객 기술을 검토할 때 쓰는 실사 방법론을 AI 모델 선정에 그대로 적용해봤습니다.

그리고 그 과정에서 기술사업화의 가장 전형적인 교훈을 다시 한번 확인했습니다.

성능표 1위와, 실제로 우리 환경에 도입할 수 있는 모델은 서로 다른 이야기였습니다.

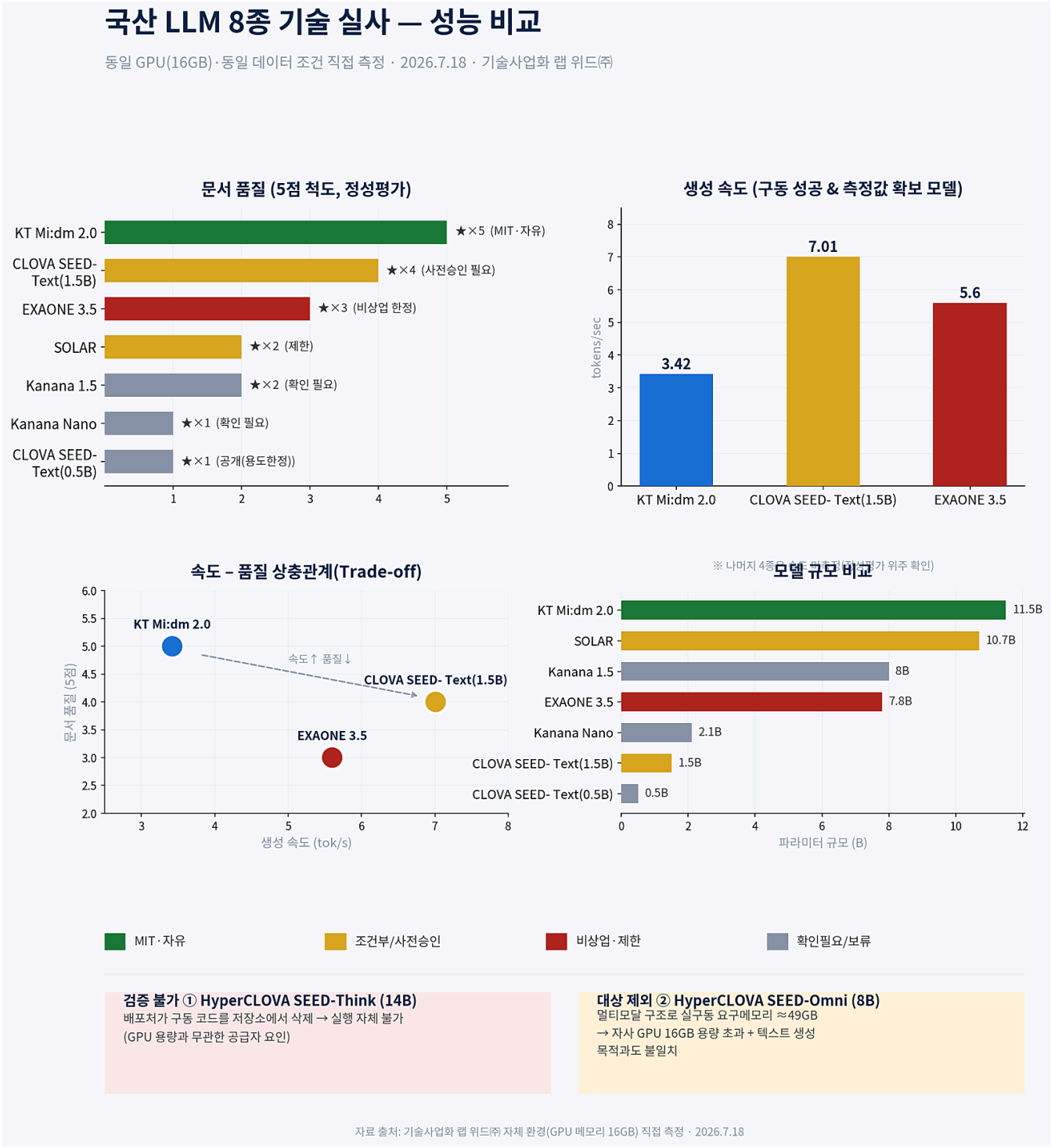
◆ 8종을 같은 조건에 세워봤습니다

국산 LLM 8종
- KT Mi:dm 2.0,
- CLOVA SEED-Text(1.5B/0.5B),
- EXAONE 3.5,
- SOLAR,
- Kanana 1.5,
- Kanana Nano,
- HyperCLOVA SEED-Think,
- SEED-Omni

상기 모델의 제조사가 공개한 벤치마크 수치가 아니라, 저희가 직접 같은 데이터로 학습시키고 같은 GPU(16GB) 환경에서 측정했습니다.

결과는 다음과 같았습니다.

▷ 실사 대상 : 8종
▷ 실제 구동 성공 : 6종
▷ 공급 문제로 검증 불가 : 2종
▷ 라이선스 제약 확인 : 3종



8종 중 2종은 성능을 측정해 볼 기회조차 없었습니다.

다만 이 두 모델의 사정은 서로 달랐습니다.

- HyperCLOVA SEED-Think(14B) : GPU 용량 문제가 아니라, 배포처가 구동 코드를 저장소에서 아예 삭제해버려 실행 자체를 시도할 수 없었습니다.

- HyperCLOVA SEED-Omni(8B) : 파라미터 규모는 8B로 다른 모델과 비슷했지만, 멀티모달(이미지·음성 처리) 구조라 실제 구동에는 약 49GB의 메모리가 필요했습니다.
저희 실험 환경(GPU 16GB)으로는 애초에 로드조차 되지 않는 크기였고, 목적(텍스트 문서 생성)에도 맞지 않아 대상에서 제외했습니다.

즉 하나는 "공급자가 코드를 내려서", 다른 하나는 "우리 환경의 VRAM 용량을 초과하는 데다 용도도 안 맞아서"
— 이유가 명확히 다릅니다.

여기에 더해 성능 상위권으로 알려진 모델 중 일부는 비상업 라이선스라 애초에 사업에 쓸 수 없는 모델이었습니다.

순위표만 보고 판단했다면, 도입을 결정하고 나서야 되돌릴 수 없는 문제와 마주쳤을 것입니다.

◆ 속도와 품질은 반비례했습니다

가장 빠른 모델(CLOVA 1.5B, 7.01 tok/s)은 가장 느린 모델(KT Mi:dm 2.0, 3.42 tok/s)보다 2배 빨랐습니다.
그러나 생성한 문서 내용은 일반론에 머물렀습니다.

반대로 속도가 가장 느렸던 KT Mi:dm 2.0은 지원 요건(매출 100억원 미만), 사업비 구조(정부 80% 이내, 과제당 3천만원), 시장 규모(세계 2조원)까지 구체적으로 채워 넣었습니다.

기술 선택은 "무엇이 더 좋은가"가 아니라 "무엇에 쓸 것인가"에서 출발해야 한다는 원칙이 AI 모델 선정에서도 그대로 확인된 셈입니다.

◆ 이번 실사로 다시 확인한 4가지 원칙

저희가 고객사의 기술이전·R&D 사전기획을 검토할 때 쓰는 기준을, 이번엔 AI 모델에 그대로 적용해봤습니다.

① 성능보다 권리관계를 먼저 본다
성능 상위권 모델 중 하나는 비상업 라이선스였습니다. 연구·테스트는 되지만 수익 활동에는 쓸 수 없습니다. 기술이전에 대입하면 — 특허의 존속기간·실시권 범위·개량기술 귀속을 확인하지 않고 사업계획부터 세우는 것과 같습니다.

② 공급 안정성은 성능만큼 중요하다
공개돼 있던 14B 모델은 배포처가 구동 코드를 삭제해 현재는 실행 자체가 불가능했습니다. 공식 발표와 실제 사용 가능성은 별개였습니다. 기술이전에 대입하면 — 기술 보유기관의 유지보수 의지, 핵심 인력 이탈 여부를 확인하지 않으면 도입 후 고립됩니다.

③ 조건을 통제하지 않은 성능 수치는 신뢰할 수 없다
초기 측정에서는 거의 모든 모델이 같은 문장을 무한 반복하는 실패를 보였습니다. 그런데 설정값 하나(반복 억제 계수)를 조정하자 같은 모델이 실무급 문서를 생성했습니다. 문제는 모델이 아니라 조건이었습니다. 실제로 이 조정만으로 순위가 바뀌었습니다.

④ 성능은 도입 환경 안에서만 의미가 있다
이번 검증은 GPU 메모리 16GB 환경에서 이뤄졌습니다. 1위 모델이 이미 13.45GB를 사용해 14B급이 사실상 상한이었습니다.
즉 이 순위는 "해당 환경에서 구동 가능한 것들 중"의 순위입니다. 경제성 평가에 대입하면 — 기술의 이론적 최대 성능이 아니라, 그 기업의 설비·인력·자금 안에서 실현 가능한 성능이 사업성을 결정합니다.

◆ 이 실사에도 한계는 있습니다

저희가 확인한 것도 절대적인 결론은 아닙니다.

- 이번 순위는 GPU 메모리 16GB라는 저희 환경 기준입니다.
더 큰 메모리를 쓸 수 있는 환경이라면 SEED-Omni 같은 대형·멀티모달 모델도 결과가 달라질 수 있습니다.

- 문서 품질 평가는 저희 팀의 정성 평가(5점 척도)를 기준으로 했습니다.
평가자의 판단이 개입될 수 있는 부분입니다.

- R&D 문서라는 단일 도메인 데이터로만 학습·측정했기 때문에, 다른 분야(법률, 의료 등) 문서에도 같은 결과가 나올지는 별도로 확인이 필요합니다.

- 대상은 2026년 7월 18일 시점 공개된 국산 sLLM 8종으로 한정했습니다.
이후 나온 신모델이나 해외 모델은 포함하지 않았습니다.

이런 한계를 감추지 않고 밝히는 것도 저희가 실사를 할 때 지키는 원칙 중 하나입니다.

◆ 저희가 일하는 방식

이번 실사에서 저희는 자료를 그대로 믿지 않고 8종을 모두 같은 조건에서 직접 측정했습니다.

그 과정에서 제조사 발표 자료만 봤다면 알 수 없었을 문제 3가지를 발견했습니다.

같은 기준을 저희는 고객사의 기술이전 타당성 검토, R&D 사전기획, 사업화 전략 수립, 사전 경제성 평가에 동일하게 적용합니다.

기술을 도입할지 판단해야 하는 상황이라면, 성능 자료를 보기 전에 권리관계·공급 안정성·도입 환경부터 함께 검토해보시길 권합니다.

기술사업화 랩 위드㈜
기술이전 · R&D 사전기획 · BM 구축 · 시장조사분석 · 사업화 전략 · 사전 경제성 평가 · 시드투자 · IR 지원

본 자료의 수치는 2026.07.18 자사 환경(GPU 메모리 16GB)에서 직접 측정한 값이며, 해당 조건에서 구동 가능한 모델을 대상으로 합니다.